Data Preparation For Data Science:

Womit Sie „DATA=“ in den analytischen Procedures von SAS am besten füttern? – Teil 2

Gerhard Svolba

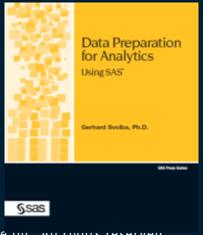
Data Scientist, SAS Austria

#datapreparation4datascience

Medium | LinkedIn | Github | SAS-Books

Youtube: DataPreparation4DataScience

Data Science Use Cases



Data Preparation for Analytics Using SAS

Gerhard Svolba, Ph.D.

SAS Press Series

Data Quality for Analytics Using SAS

Dr. Gerhard Svolba

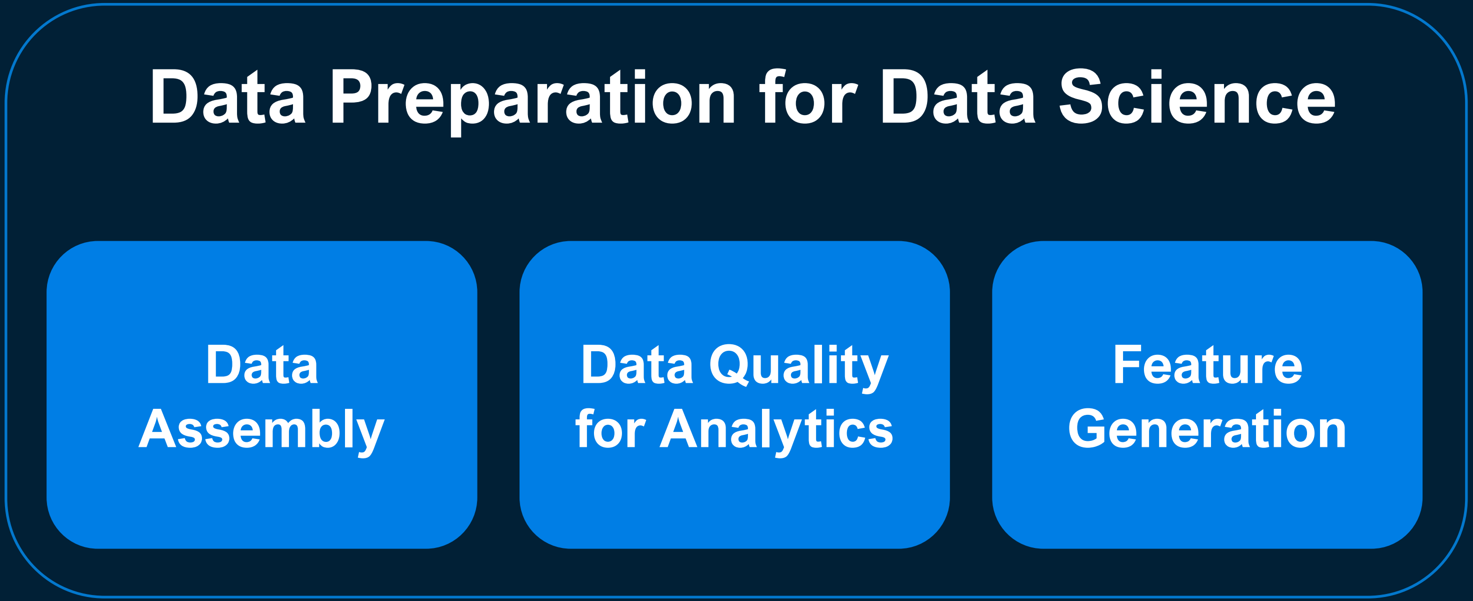
SAS Press

Applying Data Science

Business Case Studies Using SAS

Gerhard Svolba

Ein Thema mit vielen Dimensionen

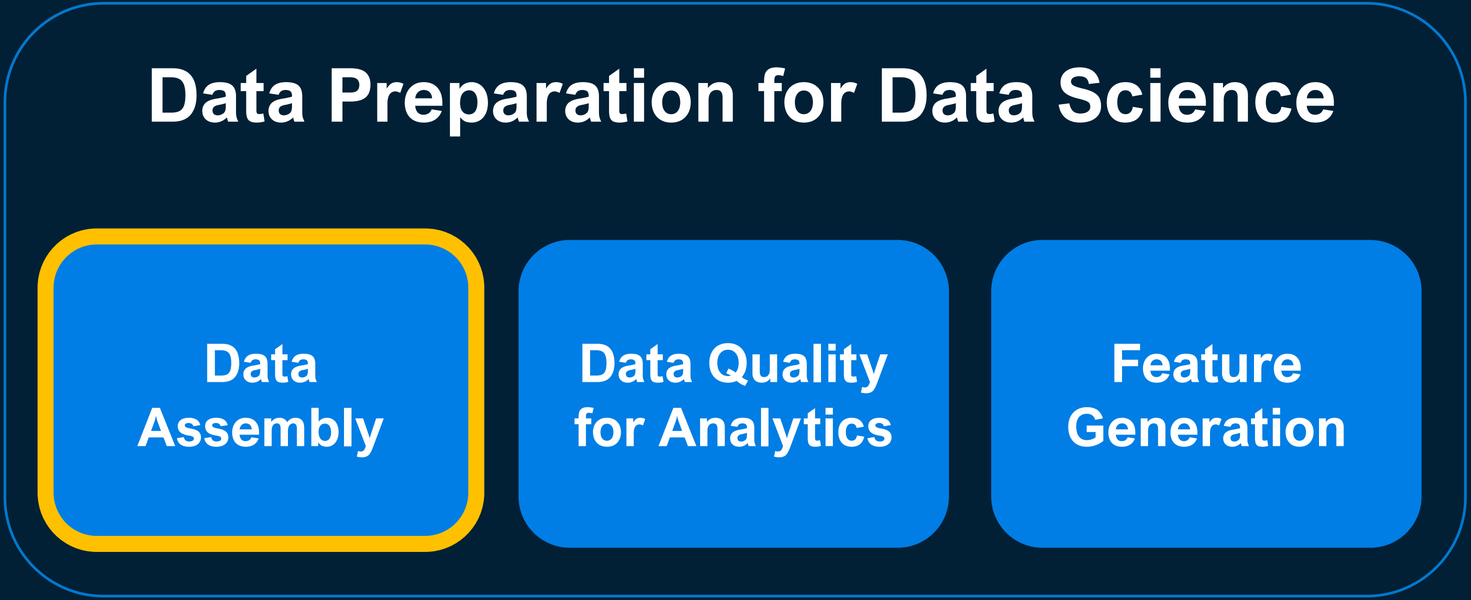


Data Preparation for Data Science

Data Assembly

Data Quality for Analytics

Feature Generation



Data Preparation for Data Science

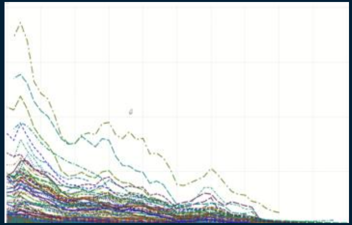
Data Assembly

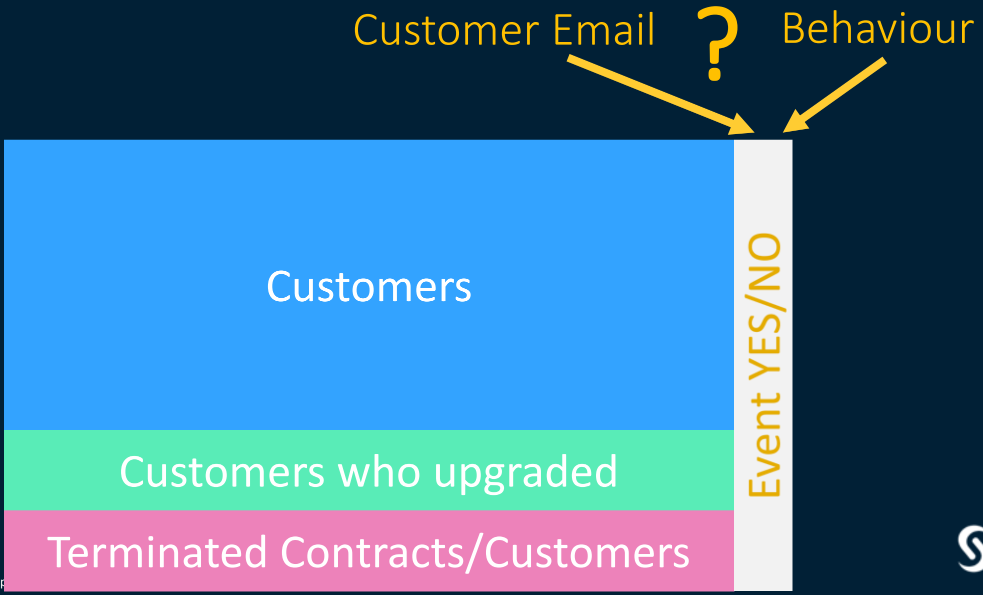
Data Quality for Analytics

Feature Generation

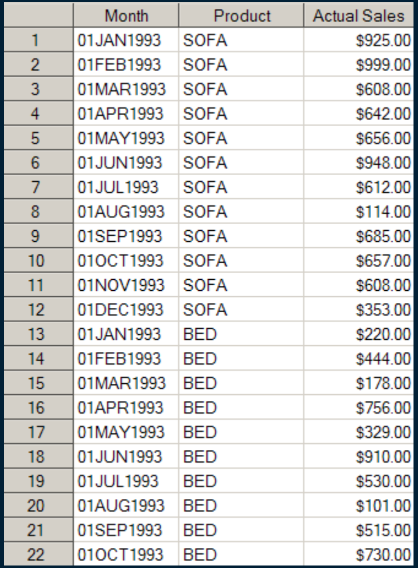
Can you build a machine learning model that predicts the cancellation risk of our customers?

- What do you mean by “cancellation”?
  - Do you mean the full cancellation of the product or a decline in usage?
  - Do you want to include customers that have canceled the product but have started to use a more (or less) advanced product?
  - Do you also want to consider customers who did not cancel themselves but were canceled by our company?



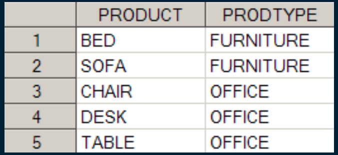


4 Methods How to Join a (Lookup) Table to a Master Table



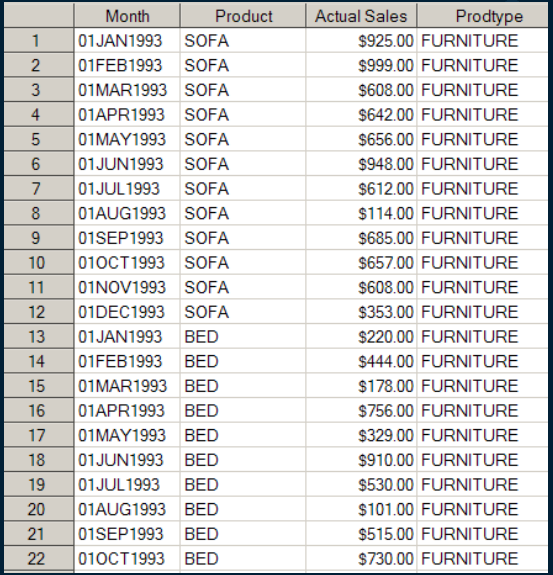
	Month	Product	Actual Sales
1	01JAN1993	SOFA	$925.00
2	01FEB1993	SOFA	$999.00
3	01MAR1993	SOFA	$608.00
4	01APR1993	SOFA	$642.00
5	01MAY1993	SOFA	$656.00
6	01JUN1993	SOFA	$948.00
7	01JUL1993	SOFA	$612.00
8	01AUG1993	SOFA	$114.00
9	01SEP1993	SOFA	$685.00
10	01OCT1993	SOFA	$657.00
11	01NOV1993	SOFA	$608.00
12	01DEC1993	SOFA	$353.00
13	01JAN1993	BED	$220.00
14	01FEB1993	BED	$444.00
15	01MAR1993	BED	$178.00
16	01APR1993	BED	$756.00
17	01MAY1993	BED	$329.00
18	01JUN1993	BED	$910.00
19	01JUL1993	BED	$530.00
20	01AUG1993	BED	$101.00
21	01SEP1993	BED	$515.00
22	01OCT1993	BED	$730.00

+



	PRODUCT	PRODTYPE
1	BED	FURNITURE
2	SOFA	FURNITURE
3	CHAIR	OFFICE
4	DESK	OFFICE
5	TABLE	OFFICE

=



	Month	Product	Actual Sales	Prodtype
1	01JAN1993	SOFA	$925.00	FURNITURE
2	01FEB1993	SOFA	$999.00	FURNITURE
3	01MAR1993	SOFA	$608.00	FURNITURE
4	01APR1993	SOFA	$642.00	FURNITURE
5	01MAY1993	SOFA	$656.00	FURNITURE
6	01JUN1993	SOFA	$948.00	FURNITURE
7	01JUL1993	SOFA	$612.00	FURNITURE
8	01AUG1993	SOFA	$114.00	FURNITURE
9	01SEP1993	SOFA	$685.00	FURNITURE
10	01OCT1993	SOFA	$657.00	FURNITURE
11	01NOV1993	SOFA	$608.00	FURNITURE
12	01DEC1993	SOFA	$353.00	FURNITURE
13	01JAN1993	BED	$220.00	FURNITURE
14	01FEB1993	BED	$444.00	FURNITURE
15	01MAR1993	BED	$178.00	FURNITURE
16	01APR1993	BED	$756.00	FURNITURE
17	01MAY1993	BED	$329.00	FURNITURE
18	01JUN1993	BED	$910.00	FURNITURE
19	01JUL1993	BED	$530.00	FURNITURE
20	01AUG1993	BED	$101.00	FURNITURE
21	01SEP1993	BED	$515.00	FURNITURE
22	01OCT1993	BED	$730.00	FURNITURE

Joining the lookup table explicitly

- Proc SQL
- Datastep

„Applying“ the lookup table to the source table

- SAS Format
- Hash Table

Method 1+2: Joining the Lookup Table Explicitly

```
PROC SQL;
 CREATE TABLE prdsale_sql_lj
 AS SELECT *
    FROM prdsale AS a
    LEFT JOIN lookup  AS b
    ON a.product = b.product
    ORDER BY product, month;
QUIT;
```

```
proc sort data = lookup;
by product;run;
proc sort data = prdsale;
by product;run;

data prdsale_ds;
 merge prdsale(in=in1)
       lookup;
 by product;
 if in1;
run;

proc sort data = prdsale_ds;
by product month;run;
```

Method 3: Using a SAS Format

```
DATA FMT_PG(RENAME =(Product=start
            ProdType=label));
 SET lookup end=last;
 RETAIN fmtname 'PG' type 'c';
RUN;

PROC FORMAT LIBRARY=work
CNTLIN=FMT_PG;
RUN;

DATA prdsale_fmt;
 SET prdsale;
 FORMAT Prodtype $12.;
 Prodtype = PUT(product,$PG.);
RUN;
```

Convert the LOOKUP Table into a control table (with specific variable names)

Use PROC FORMAT to create a SAS Format based on that table

Use the SAS Format to retrieve the value from the lookup table

Method 4: Using a Hash-Table

Define the HASH Table in the SAS Datastep

Call the HASH to retrieve the Values based on the Key-Column

```
DATA prdsale_hash;
length Product ProdType $10.;

if _n_ = 1 then do;
    declare hash h(dataset: "lookup");
    h.definekey('Product');
    h.definedata('ProdType');
    h.definedone();
    call missing(Product, ProdType);
end;

 SET prdsale;
 rc = h.find();
 drop rc;

RUN;
```

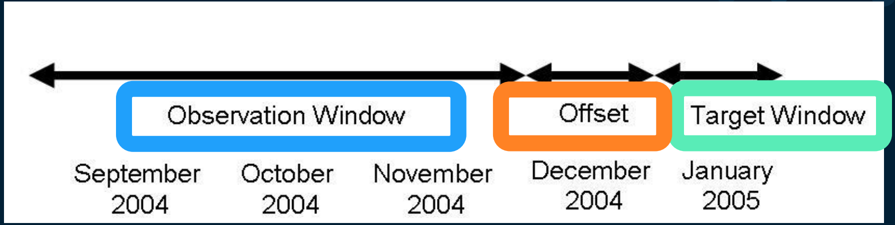
Can you build a machine learning model that predicts the cancellation risk of our customers?

- What is the business process for contacting customers?
  - What additional attributes and explanations do you need?
  - Which latency period should we consider between the availability of the scores and the execution of the marketing campaign?

Derive (additional) variable that are used for model interpretation and decisioning

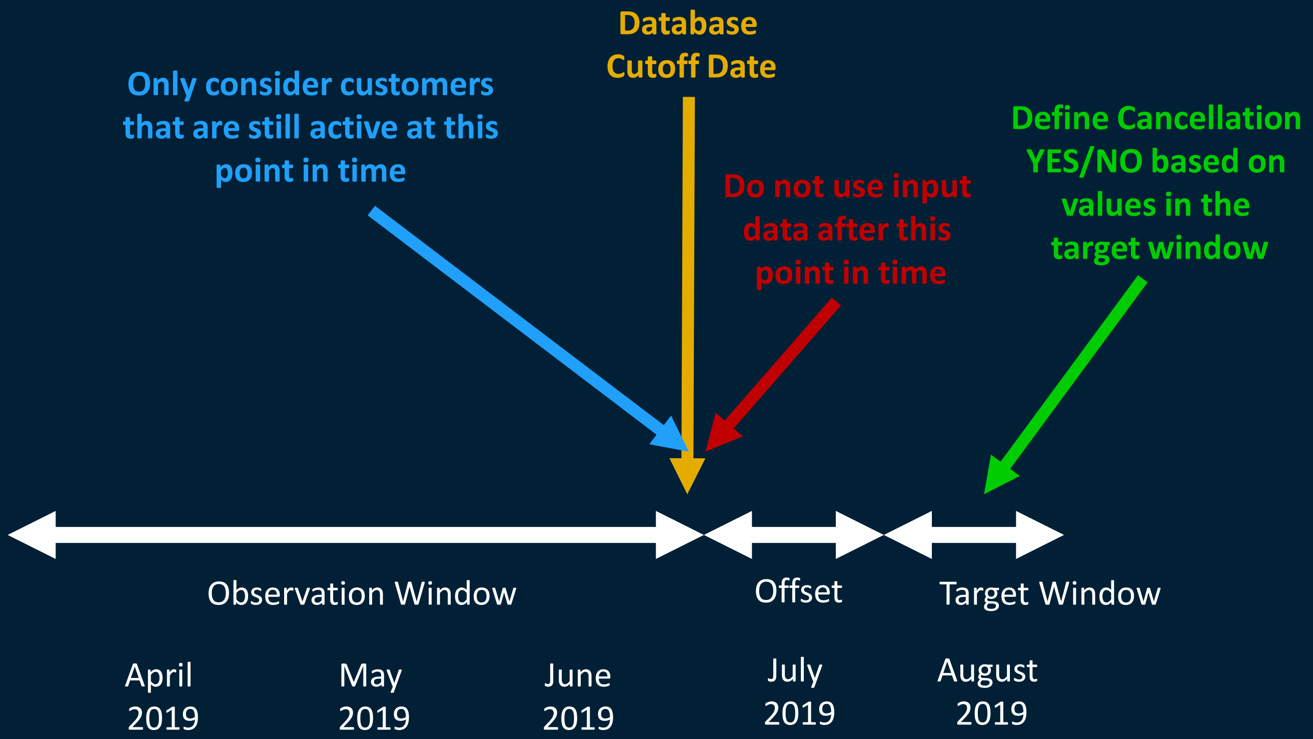
Observe the alignment on the time axis

Different Time Windows in Predictive Modeling



- Target windows: Time Interval where the target event is observed
- Example for Target Events:
  - Pay back of debt
  - Cancellation of contract
  - Purchase of product
- Observation windows: Time Interval where input data are collected
- Offset window: optional time interval between observation and target windows in order to train the model to events that occur not immediately after the snapshot date

Considerations for Supervised Machine Learning Models: Alignment on the Time Axis



Only consider customers that are still active at this point in time

Database Cutoff Date

Do not use input data after this point in time

Define Cancellation YES/NO based on values in the target window

Observation Window

Offset

Target Window

April 2019

May 2019

June 2019

July 2019

August 2019

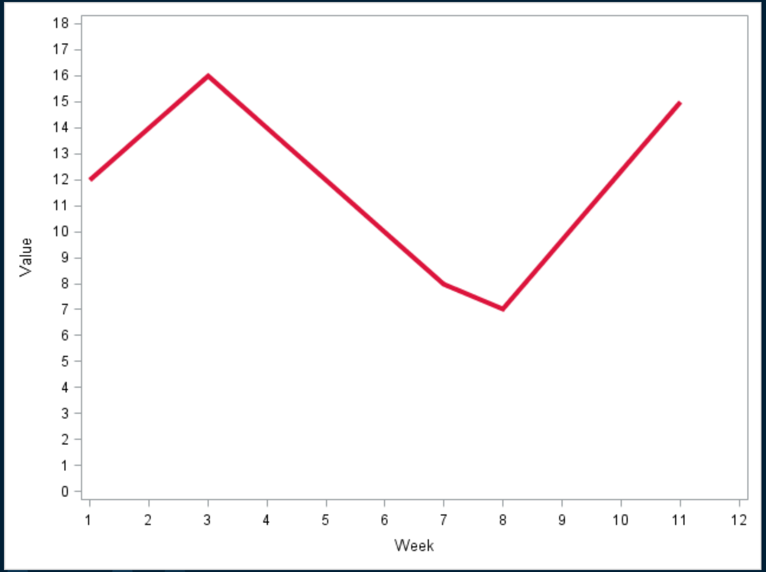
Data Preparation for Data Science

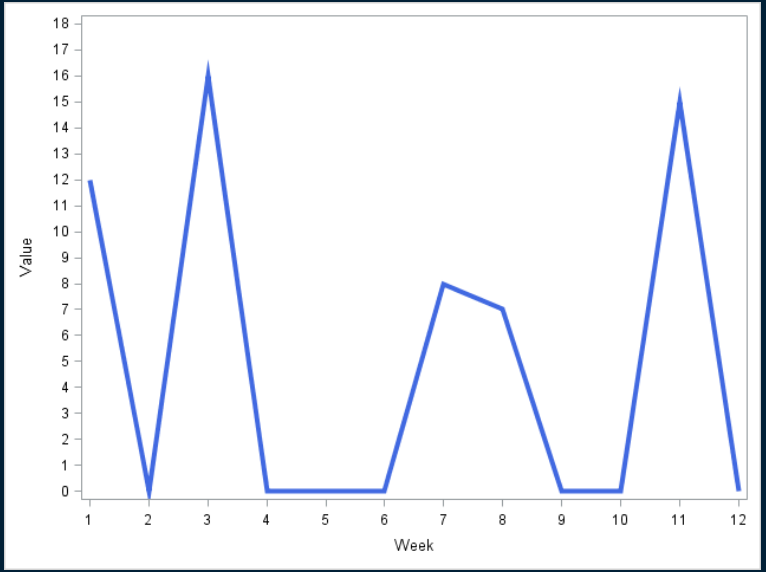
Data Assembly

Data Quality for Analytics

Feature Generation

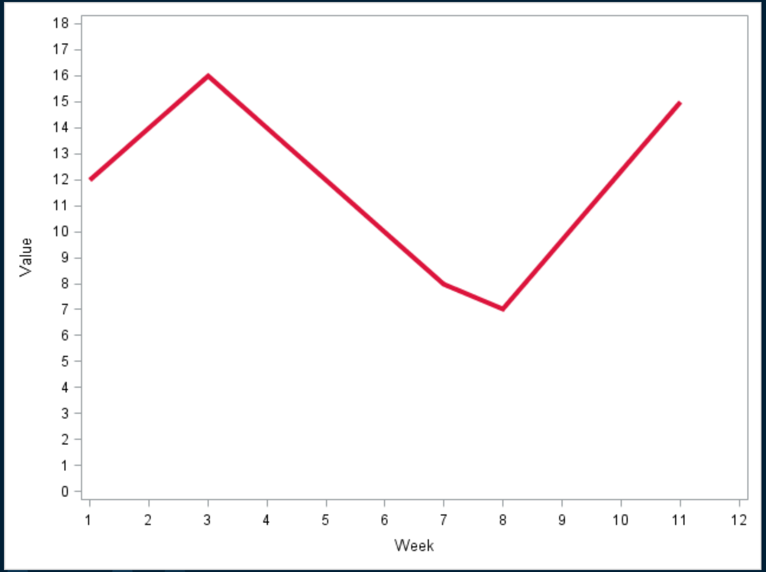
Are these two graphs based on the same data?

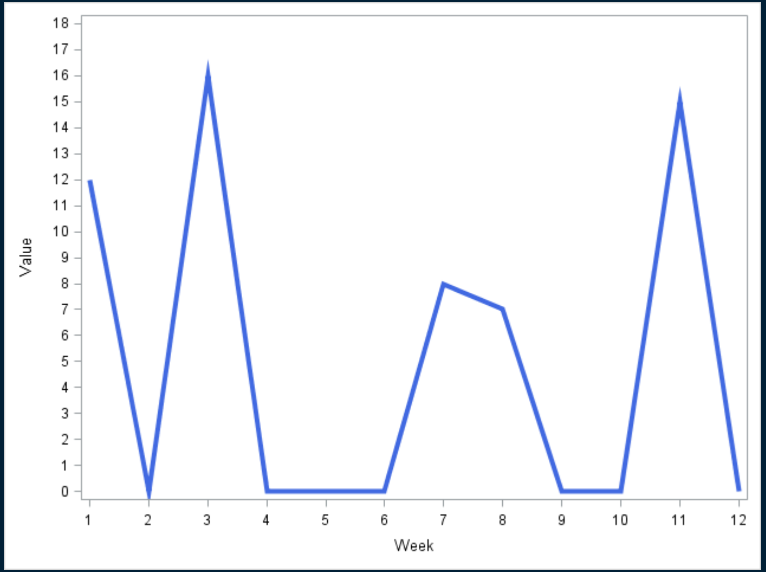




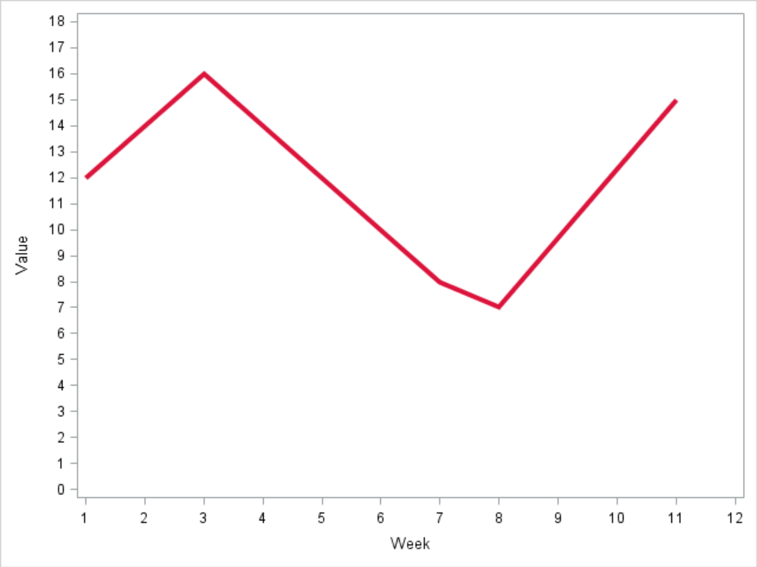
Do missing values really only matter in analytics (and not in reporting)?

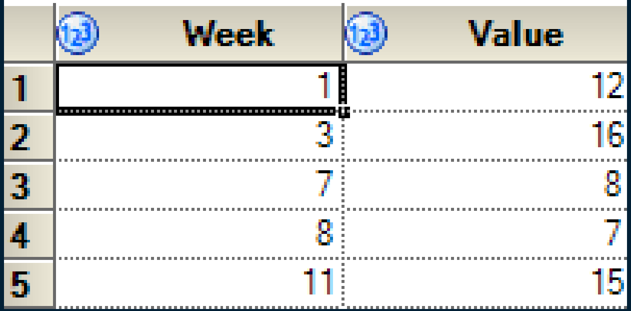
Are these two graphs based on the same data?





For some measurements (inventory data) this might be the appropriate view

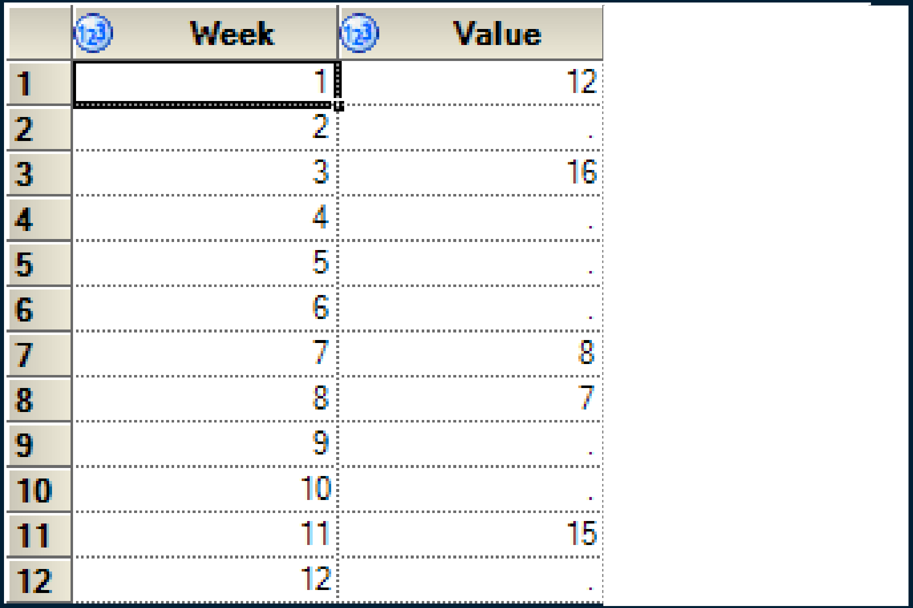




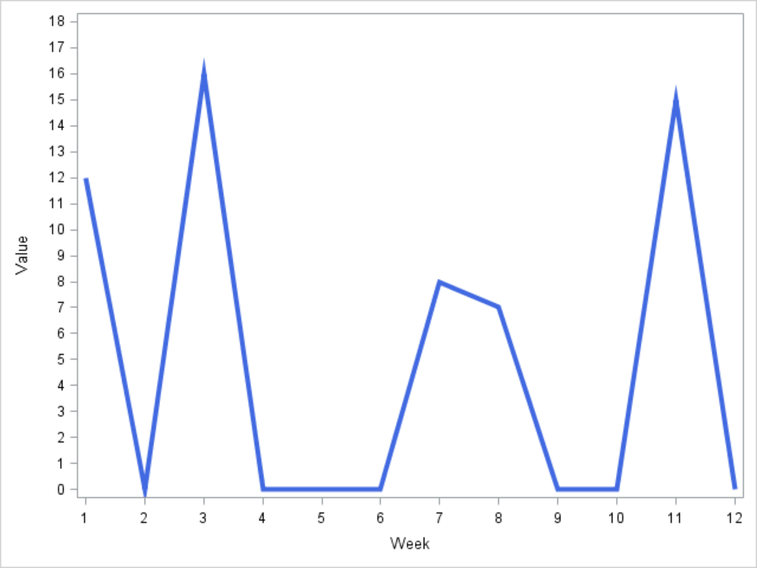
	Week	Value
1	1	12
2	3	16
3	7	8
4	8	7
5	11	15

For other measurements (movement data) this might be the appropriate view

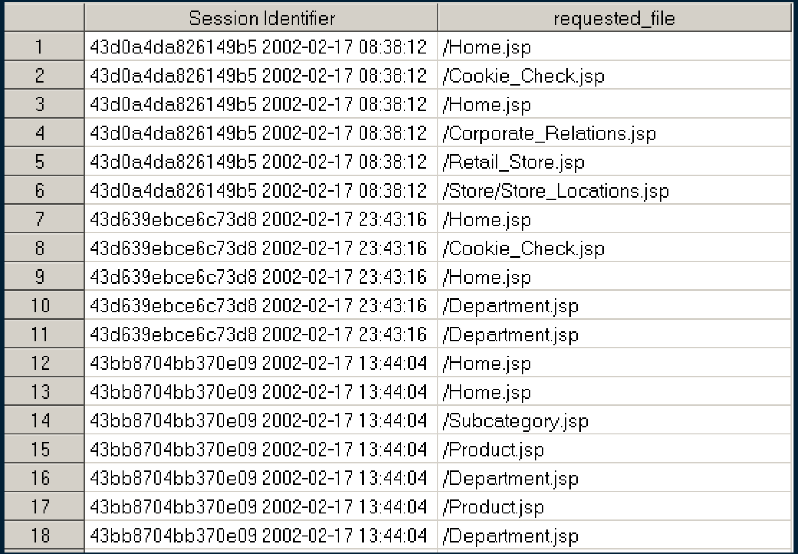
Be careful with line-charts and missing values!



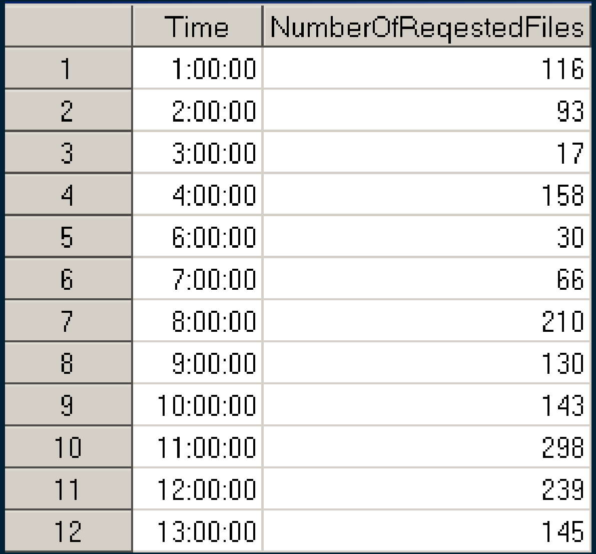
	Week	Value
1	1	12
2	2	.
3	3	16
4	4	.
5	5	.
6	6	.
7	7	8
8	8	7
9	9	.
10	10	.
11	11	15
12	12	.



Transactional Data or Timeseries Data?

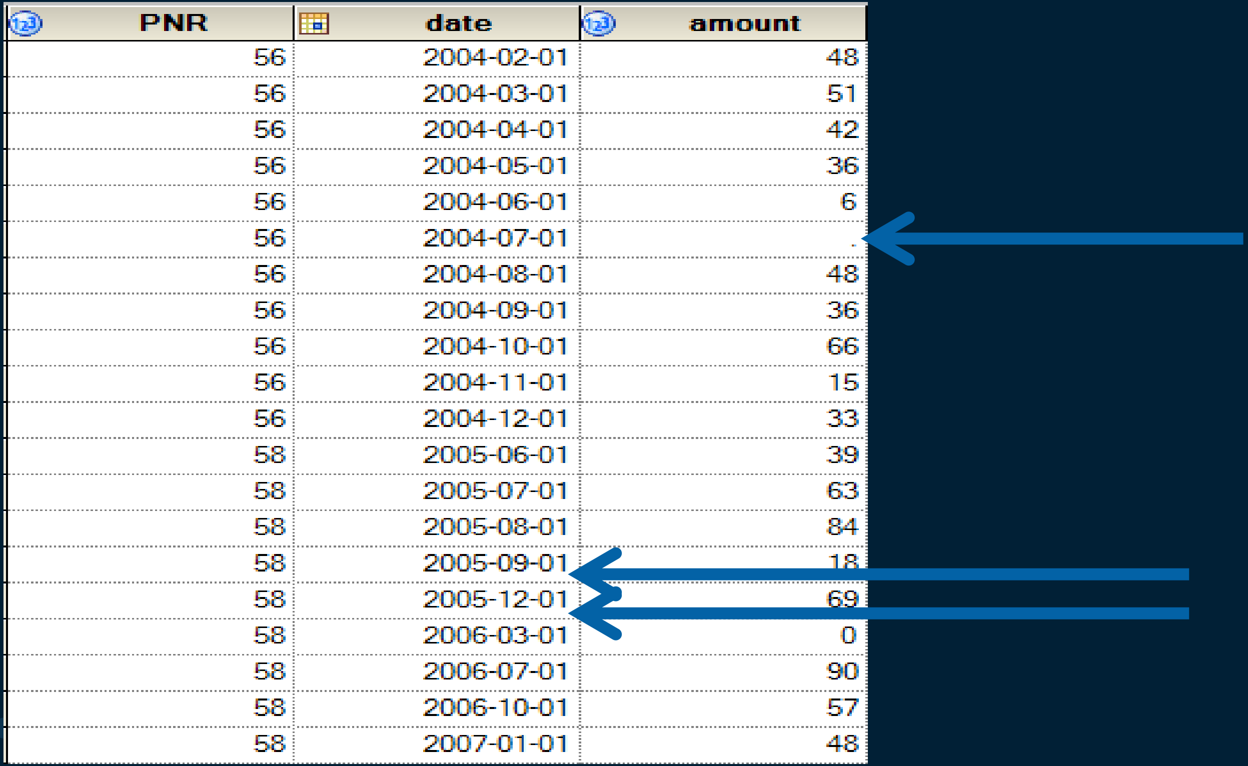


	Session Identifier	requested_file
1	43d0a4da826149b5 2002-02-17 08:38:12	/Home.jsp
2	43d0a4da826149b5 2002-02-17 08:38:12	/Cookie_Check.jsp
3	43d0a4da826149b5 2002-02-17 08:38:12	/Home.jsp
4	43d0a4da826149b5 2002-02-17 08:38:12	/Corporate_Relations.jsp
5	43d0a4da826149b5 2002-02-17 08:38:12	/Retail_Store.jsp
6	43d0a4da826149b5 2002-02-17 08:38:12	/Store/Store_Locations.jsp
7	43d639ebce6c73d8 2002-02-17 23:43:16	/Home.jsp
8	43d639ebce6c73d8 2002-02-17 23:43:16	/Cookie_Check.jsp
9	43d639ebce6c73d8 2002-02-17 23:43:16	/Home.jsp
10	43d639ebce6c73d8 2002-02-17 23:43:16	/Department.jsp
11	43d639ebce6c73d8 2002-02-17 23:43:16	/Department.jsp
12	43bb8704bb370e09 2002-02-17 13:44:04	/Home.jsp
13	43bb8704bb370e09 2002-02-17 13:44:04	/Home.jsp
14	43bb8704bb370e09 2002-02-17 13:44:04	/Subcategory.jsp
15	43bb8704bb370e09 2002-02-17 13:44:04	/Product.jsp
16	43bb8704bb370e09 2002-02-17 13:44:04	/Department.jsp
17	43bb8704bb370e09 2002-02-17 13:44:04	/Product.jsp
18	43bb8704bb370e09 2002-02-17 13:44:04	/Department.jsp



	Time	NumberOfReqestedFiles
1	1:00:00	116
2	2:00:00	93
3	3:00:00	17
4	4:00:00	158
5	6:00:00	30
6	7:00:00	66
7	8:00:00	210
8	9:00:00	130
9	10:00:00	143
10	11:00:00	298
11	12:00:00	239
12	13:00:00	145

Explicit or implicit missing values in longitudinal data



PNR	date	amount
56	2004-02-01	48
56	2004-03-01	51
56	2004-04-01	42
56	2004-05-01	36
56	2004-06-01	6
56	2004-07-01	.
56	2004-08-01	48
56	2004-09-01	36
56	2004-10-01	66
56	2004-11-01	15
56	2004-12-01	33
58	2005-06-01	39
58	2005-07-01	63
58	2005-08-01	84
58	2005-09-01	18
58	2005-12-01	69
58	2006-03-01	0
58	2006-07-01	90
58	2006-10-01	57
58	2007-01-01	48

Existing Record
Value Missing

Missing Record
No Continuity

Replacing and interpolating missing values in longitudinal data with SAS

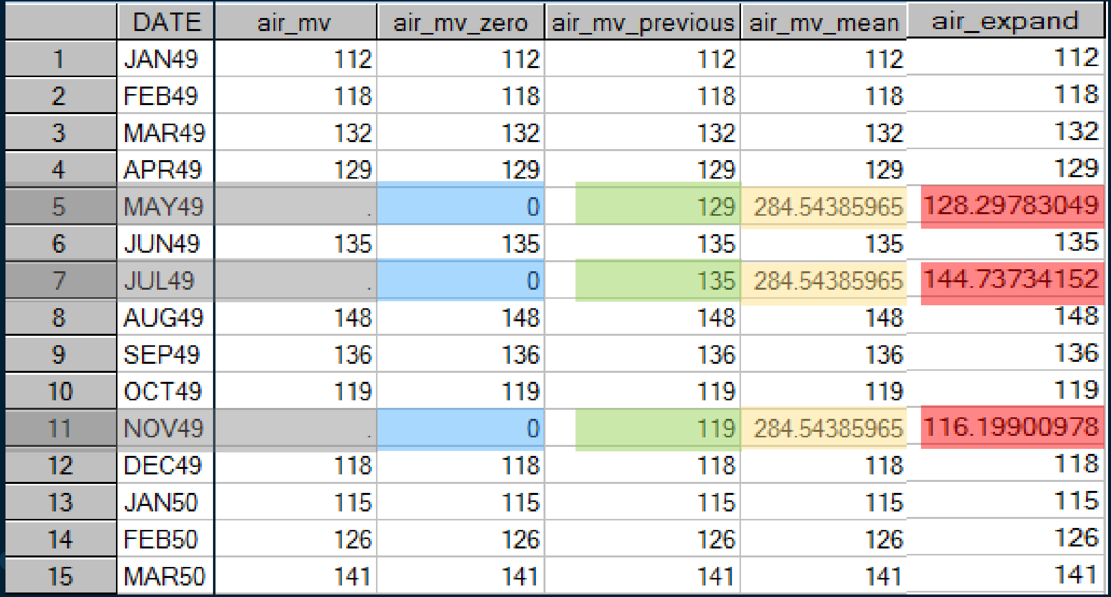
Insert missing records

Replace with 0

Replace with last known value

Replace with mean

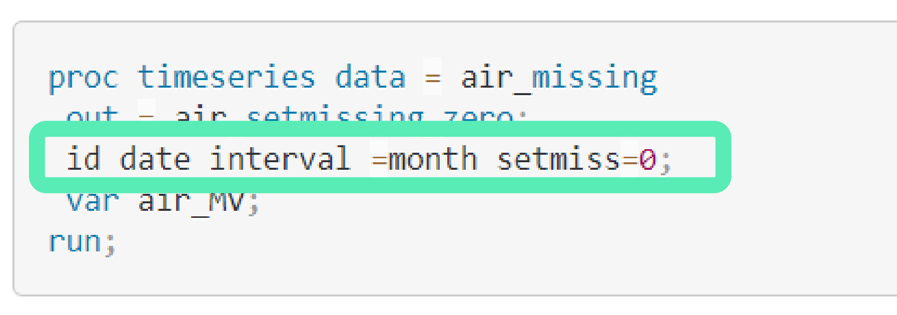
Interpolate based on splines



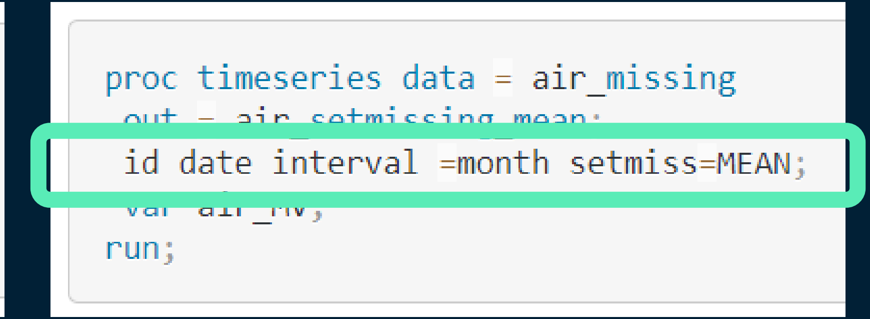
	DATE	air_mv	air_mv_zero	air_mv_previous	air_mv_mean	air_expand
1	JAN49	112	112	112	112	112
2	FEB49	118	118	118	118	118
3	MAR49	132	132	132	132	132
4	APR49	129	129	129	129	129
5	MAY49	.	0	129	284.54385965	128.29783049
6	JUN49	135	135	135	135	135
7	JUL49	.	0	135	284.54385965	144.73734152
8	AUG49	148	148	148	148	148
9	SEP49	136	136	136	136	136
10	OCT49	119	119	119	119	119
11	NOV49	.	0	119	284.54385965	116.19900978
12	DEC49	118	118	118	118	118
13	JAN50	115	115	115	115	115
14	FEB50	126	126	126	126	126
15	MAR50	141	141	141	141	141

Use PROC TIMESERIES and PROC EXPAND for these tasks!

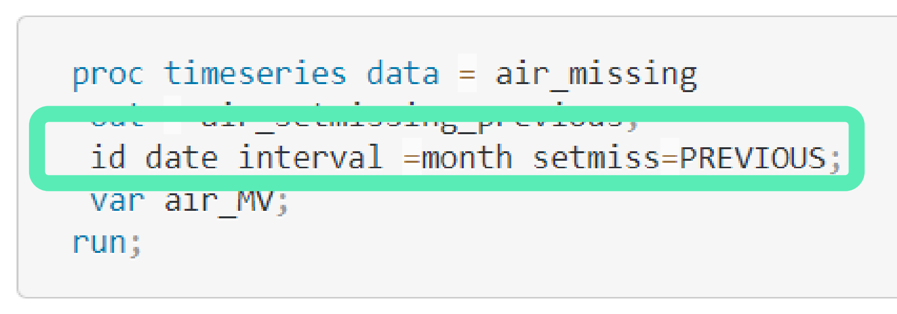
Aggregation and Processing of Data in One Step with the TIMESERIES Procedure



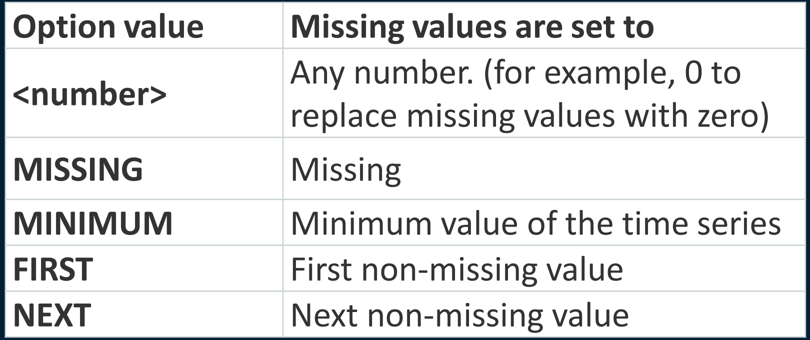
```
proc timeseries data = air_missing
 out = air_setmissing_zero;
 id date interval =month setmiss=0;
 var air_MV;
run;
```



```
proc timeseries data = air_missing
 out = air_setmissing_mean;
 id date interval =month setmiss=MEAN;
 var air_MV;
run;
```



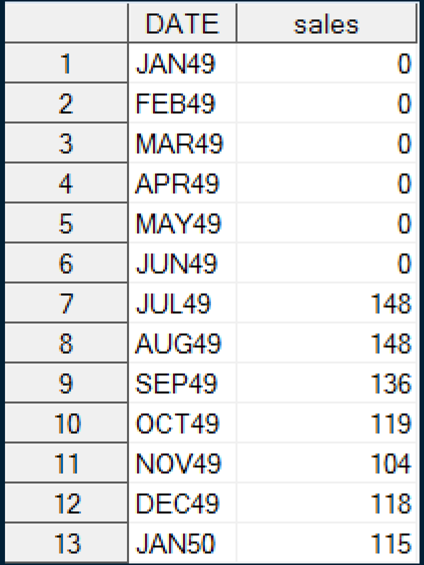
```
proc timeseries data = air_missing
 out = air_setmissing_previous;
 id date interval =month setmiss=PREVIOUS;
 var air_MV;
run;
```



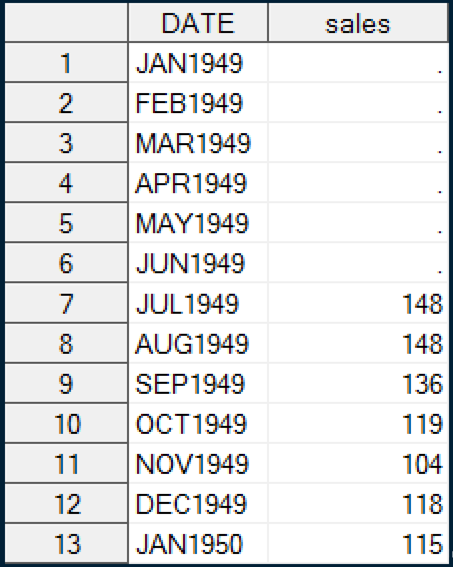
Option value	Missing values are set to
<number>	Any number. (for example, 0 to replace missing values with zero)
MISSING	Missing
MINIMUM	Minimum value of the time series
FIRST	First non-missing value
NEXT	Next non-missing value

Convert Leading and Trailing Zeros to Missing Values

```
proc timeseries
    data=sales_original
    out=sales corrected;
id date interval=month
        zeromiss=both;
var sales;
run;
```

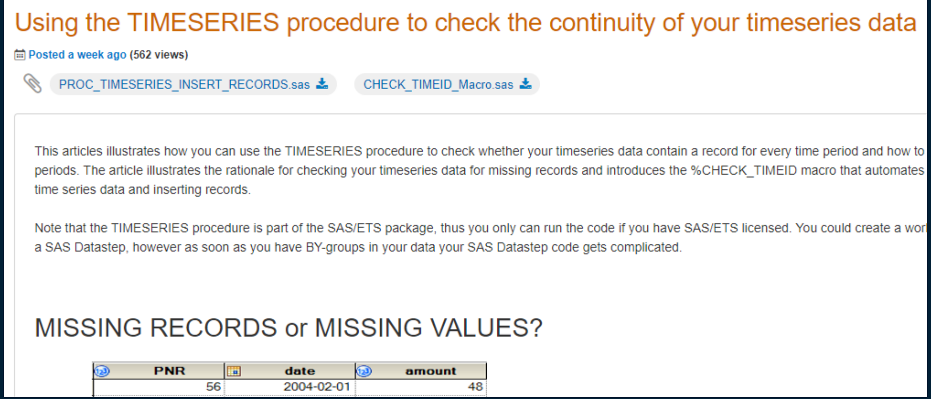


	DATE	sales
1	JAN49	0
2	FEB49	0
3	MAR49	0
4	APR49	0
5	MAY49	0
6	JUN49	0
7	JUL49	148
8	AUG49	148
9	SEP49	136
10	OCT49	119
11	NOV49	104
12	DEC49	118
13	JAN50	115



	DATE	sales
1	JAN1949	.
2	FEB1949	.
3	MAR1949	.
4	APR1949	.
5	MAY1949	.
6	JUN1949	.
7	JUL1949	148
8	AUG1949	148
9	SEP1949	136
10	OCT1949	119
11	NOV1949	104
12	DEC1949	118
13	JAN1950	115

Two related Articles at Communities.sas.com



Using the TIMESERIES procedure to check the continuity of your timeseries data

Posted a week ago (562 views)

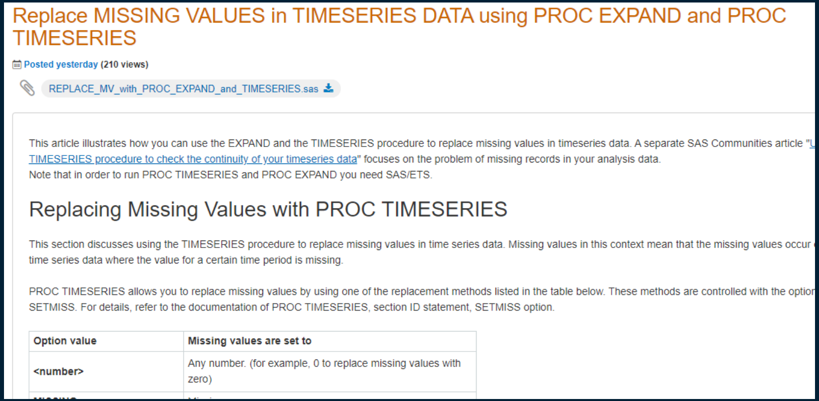
PROC_TIMESERIES_INSERT_RECORDS.sas   CHECK_TIMEID_Macro.sas

This articles illustrates how you can use the TIMESERIES procedure to check whether your timeseries data contain a record for every time period and how to periods. The article illustrates the rationale for checking your timeseries data for missing records and introduces the %CHECK_TIMEID macro that automates time series data and inserting records.

Note that the TIMESERIES procedure is part of the SAS/ETS package, thus you only can run the code if you have SAS/ETS licensed. You could create a wor a SAS Datastep, however as soon as you have BY-groups in your data your SAS Datastep code gets complicated.

MISSING RECORDS or MISSING VALUES?

PNR	date	amount
56	2004-02-01	48



Replace MISSING VALUES in TIMESERIES DATA using PROC EXPAND and PROC TIMESERIES

Posted yesterday (210 views)

REPLACE_MV_with_PROC_EXPAND_and_TIMESERIES.sas

This article illustrates how you can use the EXPAND and the TIMESERIES procedure to replace missing values in timeseries data. A separate SAS Communities article "U TIMESERIES procedure to check the continuity of your timeseries data" focuses on the problem of missing records in your analysis data.
Note that in order to run PROC TIMESERIES and PROC EXPAND you need SAS/ETS.

Replacing Missing Values with PROC TIMESERIES

This section discusses using the TIMESERIES procedure to replace missing values in time series data. Missing values in this context mean that the missing values occur time series data where the value for a certain time period is missing.

PROC TIMESERIES allows you to replace missing values by using one of the replacement methods listed in the table below. These methods are controlled with the optior SETMISS. For details, refer to the documentation of PROC TIMESERIES, section ID statement, SETMISS option.

Option value	Missing values are set to
<number>	Any number. (for example, 0 to replace missing values with zero)

https://communities.sas.com/t5/SAS-Communities-Library/Using-the-TIMESERIES-procedure-to-check-the-continuity-of-your/ta-p/714678

https://communities.sas.com/t5/SAS-Communities-Library/Replace-MISSING-VALUES-in-TIMESERIES-DATA-using-PROC-EXPAND-and/ta-p/714806

SGF-Paper: Want an Early Picture of the Data Quality Status of Your Analysis Data? SAS® Visual Analytics Shows You How

Data Preparation for Data Science

Data Assembly

Data Quality for Analytics

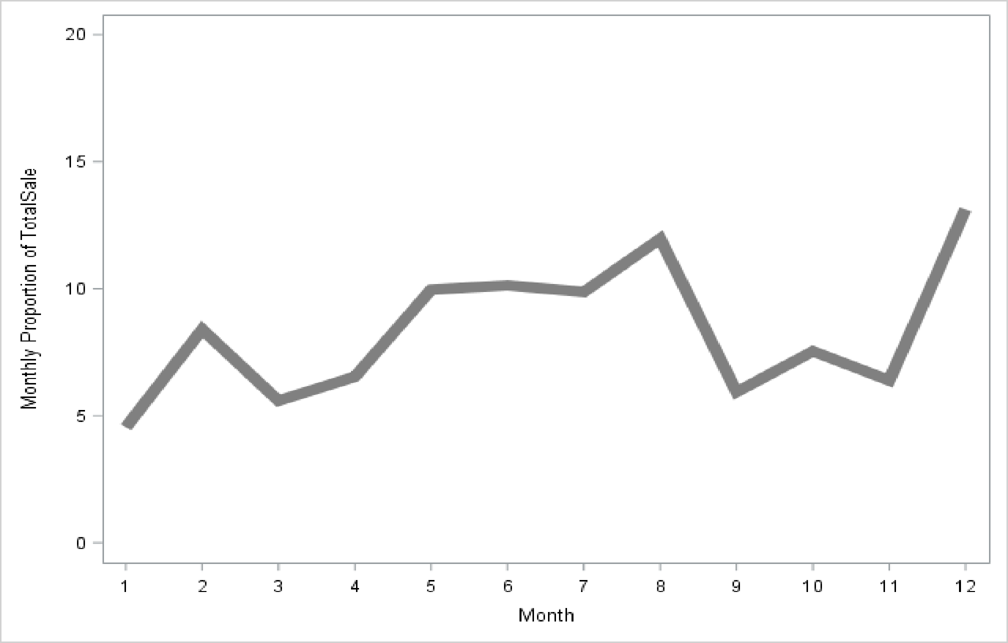
Feature Generation

Which of my sales representatives do not follow pre-defined pattern?

The demand for sub-contractors for a company in the catering business varies over the calendar year.

Sales Persons are forced to close such sub-contracts following the seasonal demand pattern.

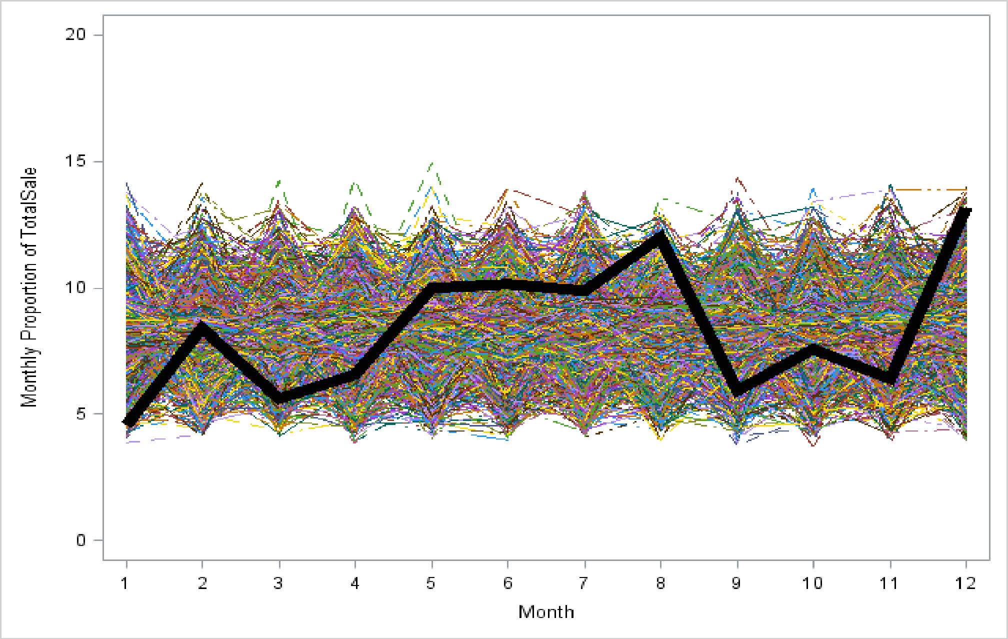
```
proc freq data=sales_historic;
table month / nocum
out=HistoricDemand(rename=(percent=HistoricPct));
weight Sales_EUR;
run;
```



Looking at the individual seasonal pattern per sales person does not help

No clear picture.

Infeasible to review all individual lines manually.

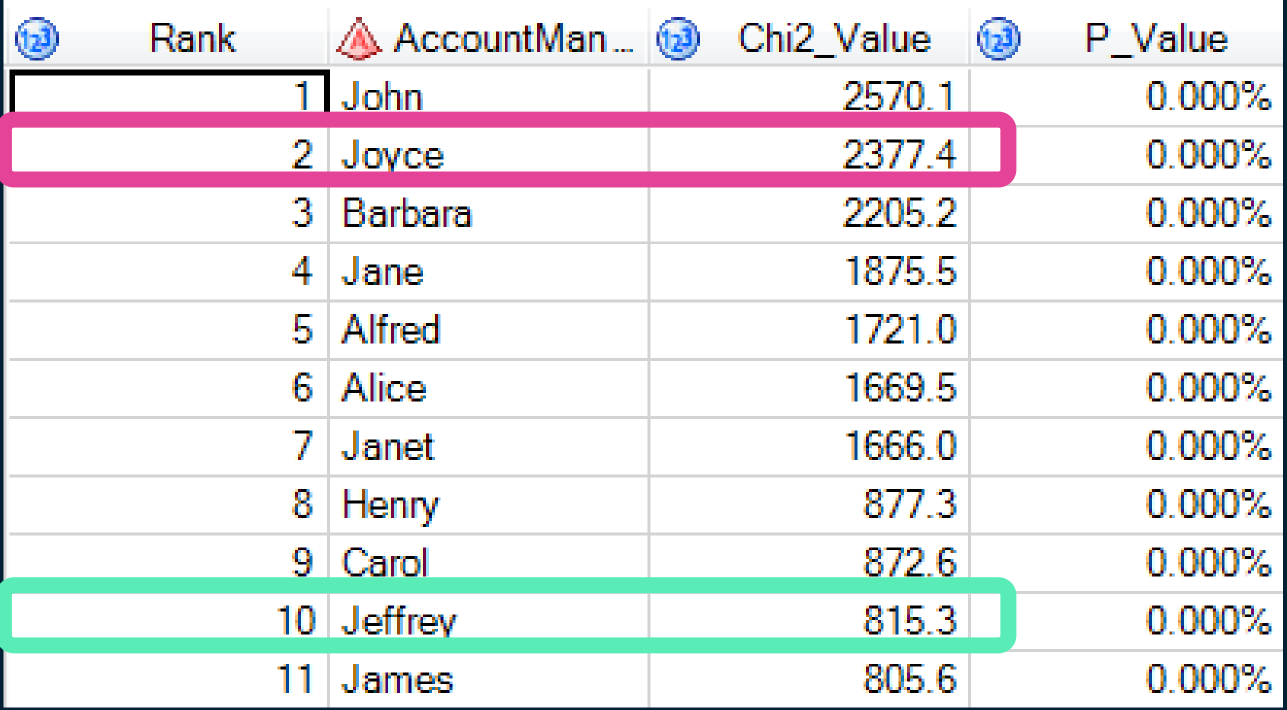


Performing a Chi2-Test Using the FREQ Procedure

```
proc freq data=sales_month;
by AccountManager;
table month / nocum out=Sales_AccMgr
chisq(testp=HistoricDemand(rename=(HistoricPct= testp )));
weight Sales_EUR;
ods output OneWayChiSq=Chi2_AccMgr(drop=table label cvalue);
run;
```

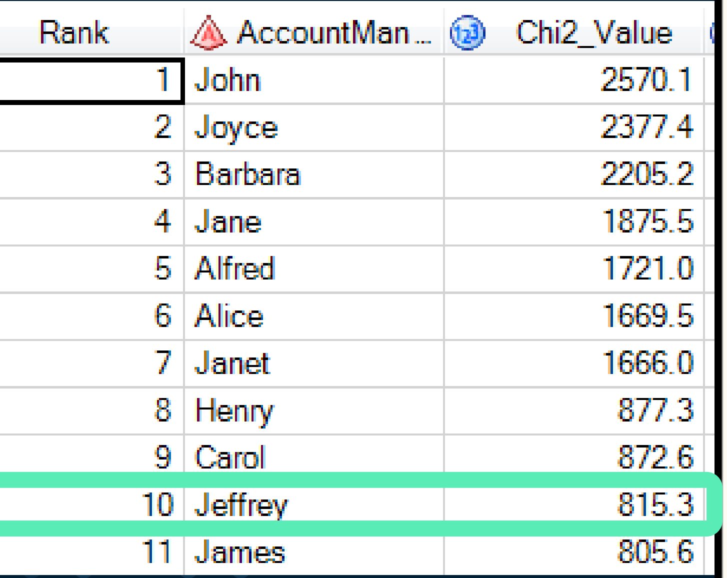
Receiving a KPI to rank analysis subjects based on their “Accordance” with the predefined pattern

(after transposing and preparing the data – see link section)

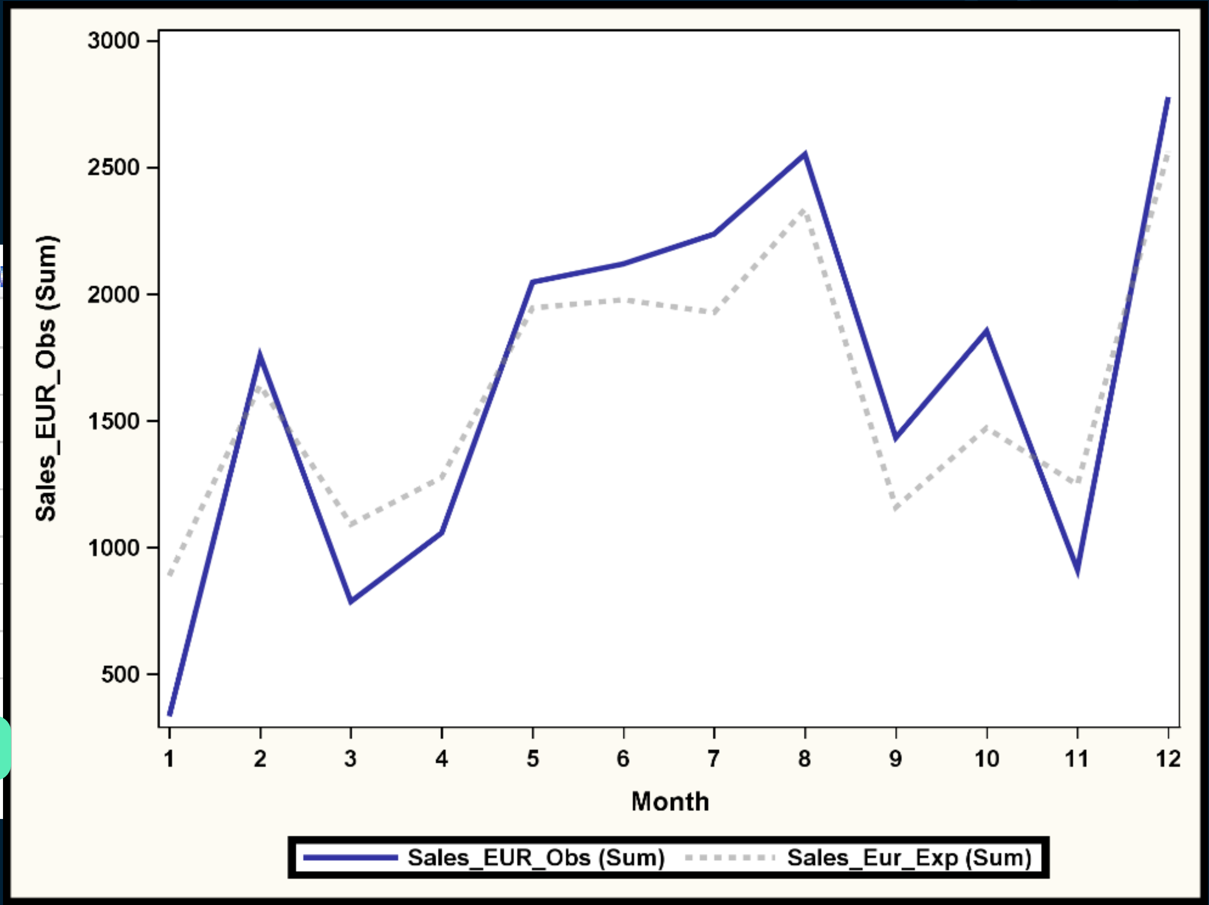


Rank	AccountMan...	Chi2_Value	P_Value
1	John	2570.1	0.000%
2	Joyce	2377.4	0.000%
3	Barbara	2205.2	0.000%
4	Jane	1875.5	0.000%
5	Alfred	1721.0	0.000%
6	Alice	1669.5	0.000%
7	Janet	1666.0	0.000%
8	Henry	877.3	0.000%
9	Carol	872.6	0.000%
10	Jeffrey	815.3	0.000%
11	James	805.6	0.000%

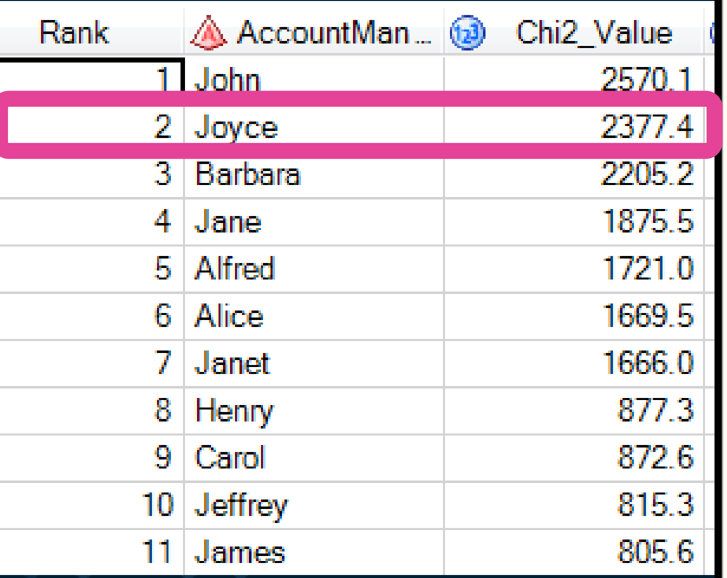
Line Chart for Jeffrey



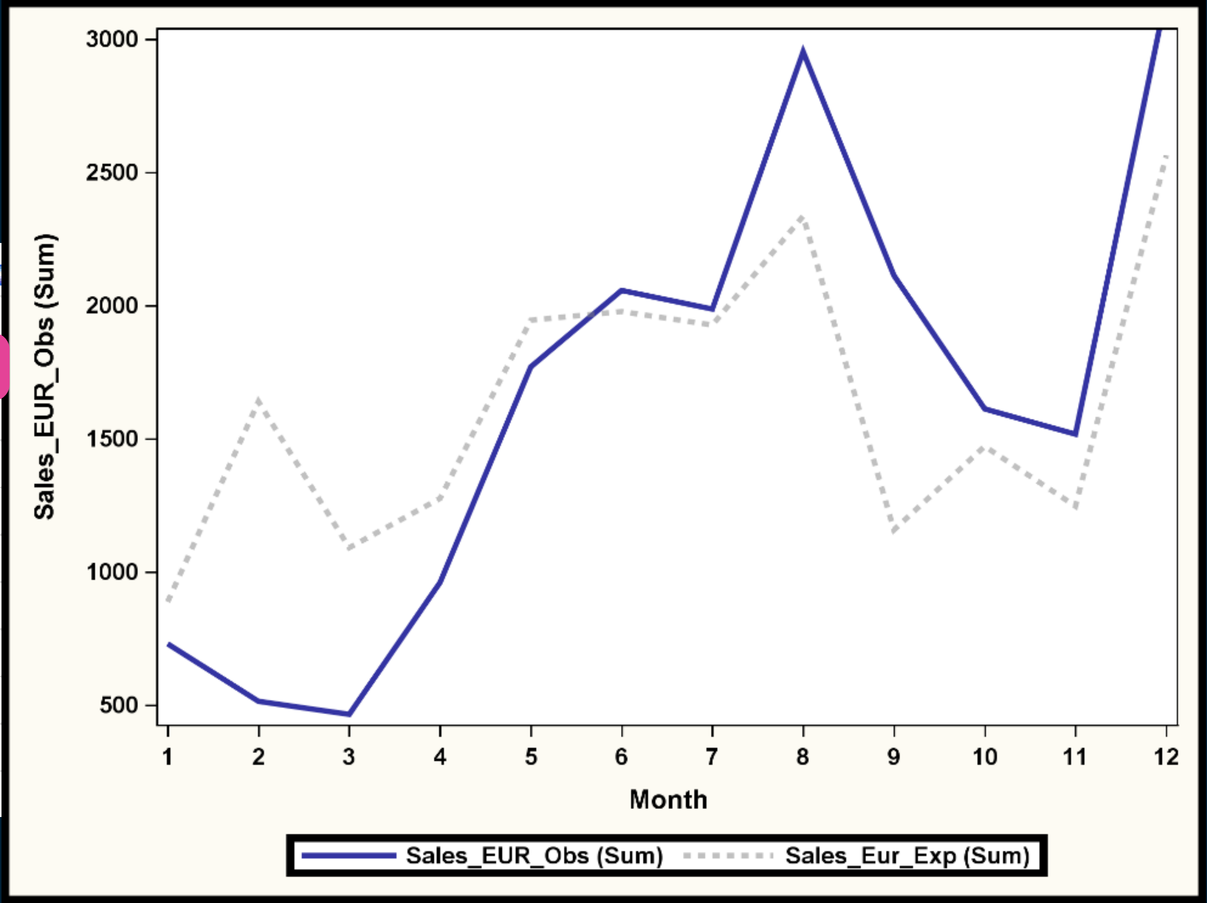
Rank	AccountMan...	Chi2_Value
1	John	2570.1
2	Joyce	2377.4
3	Barbara	2205.2
4	Jane	1875.5
5	Alfred	1721.0
6	Alice	1669.5
7	Janet	1666.0
8	Henry	877.3
9	Carol	872.6
10	Jeffrey	815.3
11	James	805.6



Line Chart for Joyce



Rank	AccountMan...	Chi2_Value
1	John	2570.1
2	Joyce	2377.4
3	Barbara	2205.2
4	Jane	1875.5
5	Alfred	1721.0
6	Alice	1669.5
7	Janet	1666.0
8	Henry	877.3
9	Carol	872.6
10	Jeffrey	815.3
11	James	805.6



Links

- Webinar at Youtube:
  Use Data Science Methods to check the Alignment of your processes with Predefined Pattern
  https://www.youtube.com/watch?v=YWqgPeVWpUg&list=PLdMxv2SumIKs0A2cQLeXg1xb9OVE8e2Yq&index=7&t=0

- SAS Programs: Github Link, Chapter 18-20
  https://github.com/gerhard1050/Applying-Data-Science-Using-SAS

Feature Engineering – Be creative!

Multiple Observation per Analysis Subject

ID	Month	Type	Billing	Usage	…
1					
1					
1					
2					
2					
3					
3					
3					
4					
4					
4					
4					

Aggregate, Transpose
Describe Behaviour

Billing_Sum	Billing_Mean	Usage_Sum	Usage_Trend	Usage_Variab	N_Trx

Interval Data

- Correlation of Values
- Course over Time
- Concentration of Values
- Seasonal Pattern

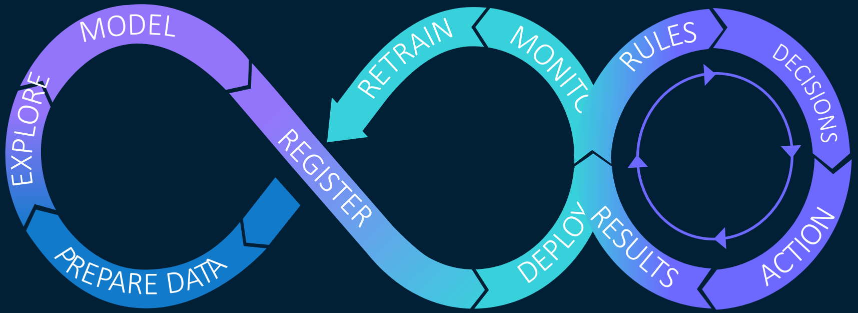
Categorical Data

- Frequency Counts
- Concatenated Frequencies
- Total and Distinct Counts

- Network Data
- Textual Data
- Images and Videos
- …

Conclusion

- Data Preparation is all over the analytic lifecycle!



- Data Preparation is much more than just coding!

All you need to prepare your data for data science is available in the integrated SAS Viya platform

- Data Preparation / Data Quality / Feature Engineering / Variety of Analytical Methods / Visualizing Relationships / Comparing Models / What-If Scenarios / Access for different Persona Roles / Model Ops / ...

Data Preparation for Data Science

Data Assembly | Data Quality for Analytics | Feature Generation

Gerhard Svolba,
Data Scientist @SAS
mailto: gerhard.svolba@sas.com

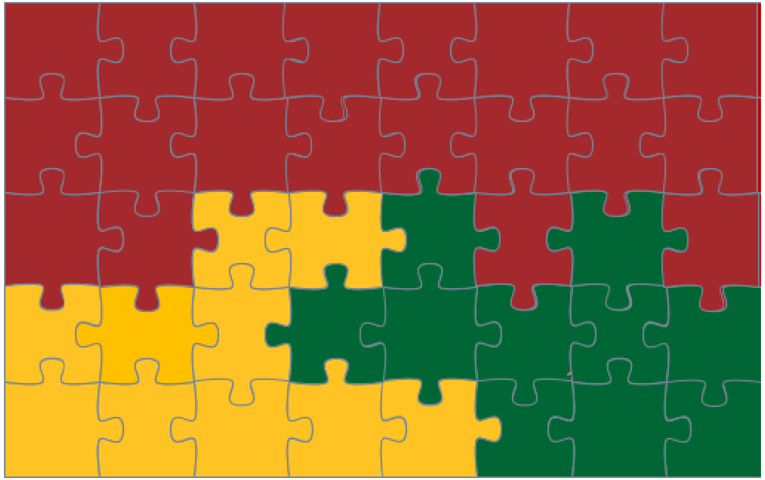
Medium | LinkedIn | Github | SAS-Books
Youtube: DataPreparation4DataScience
Data Science Use Cases

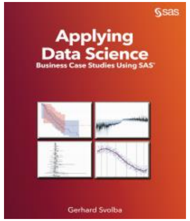
Articles and Blogs Medium LinkedIn

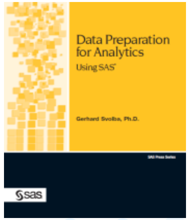
Webinars YouTube

Tipps & Tricks sas COMMUNITIES

Macros & Downloads







sas

Weitere Links

- Name: Webinar „Data Preparation for Data Science" im SAS DACH Youtube Channel
- URL: https://www.youtube.com/playlist?list=PLdMxv2SumIKsqedLBq0t_a2_6d7jZ6Akq
-
- Name: Data Preparation for Analytics Using SAS
- URL: https://github.com/gerhard1050/Data-Preparation-for-Data-Science-Using-SAS/blob/master/README.md
-
- Name: Data Quality for Analytics Using SAS
- URL: https://github.com/gerhard1050/Data-Quality-for-Data-Science-Using-SAS/blob/master/README.md
-
- Name: Applying Data Science – Business Analyses Using SAS
- URL: https://github.com/gerhard1050/Applying-Data-Science-Using-SAS/blob/master/README.md